



# A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI

Xiang Yue\*†, Yuansheng Ni\*, Kai Zhang\*, Tianyu Zheng\*,

Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun\*, Yu Su\*†, Wenhuchen Chen\*

## MMMU Team

\*Core Contributors

†Corresponding to: [xiangyue.work@gmail.com](mailto:xiangyue.work@gmail.com), [su.809@osu.edu](mailto:su.809@osu.edu), [wenhuchen@uwaterloo.ca](mailto:wenhuchen@uwaterloo.ca)

arXiv

MMMU-Pro

MMMU

Code

Leaderboard

EvalAI

Twitter

Examples

Comprehensive Disciplines	Heterogeneous Image Types	Interleaved Text and Images	Expert-level Skills Test
Engineering (26%)  Art & Design (11%)  Business (14%)  Science (23%)  Humanities & Social Sci. (9%)  Medicine (17%) 	Diagrams, Tables, Plots and Charts, Photographs, Chemical Structures, Paintings, Medical Images, Sheet Music, Geometric, Pathology images, Microscopic Images, Comics, ...	<b>Question:</b> You are shown subtraction <image 1>, T2 weighted <image 2> and T1 weighted axial <image 3> from a screening breast MRI. What is the etiology of the finding in the left breast?  <image 1> <image 2> <image 3>	<b>Perception</b>  <b>Knowledge</b> → <b>Reasoning</b> Domain Expertise, World, Linguistic, Visual Knowledge,... Logical, Spatial Commonsense, Mathematical, ...

Overview of the MMMU dataset. MMMU presents four challenges: 1) **comprehensiveness**: 11.5K college-level problems across six broad disciplines and 30 college subjects; 2) highly **heterogeneous** image types; 3) **interleaved** text and images; 4) **expert-level** perception and reasoning rooted in deep subject knowledge.

## News

🔥 [2024-09-05] Introducing [MMMU-Pro](#), a robust version of MMMU benchmark for multimodal AI evaluation!



🚀 [2024-01-31]: We added Human Expert performance on the [Leaderboard](#)! 🌟

🔥 [2023-12-04]: Our evaluation server for the test set is now available on [EvalAI](#). We welcome all submissions and look forward to your participation! 😊

## Introduction

We introduce MMMU: a new benchmark designed to evaluate multimodal models on massive multi-discipline tasks demanding college-level subject knowledge and deliberate reasoning. MMMU includes **11.5K** meticulously collected multimodal questions from college exams, quizzes, and textbooks, covering six core disciplines: Art & Design, Business, Science, Health & Medicine, Humanities & Social Science, and Tech & Engineering. These questions span **30** subjects and **183** subfields, comprising 30 highly heterogeneous image types, such as charts, diagrams, maps, tables, music sheets, and chemical structures. Unlike existing benchmarks, MMMU focuses on advanced perception and reasoning with domain-specific knowledge, challenging models to perform tasks akin to those faced by experts. Our evaluation of 14 open-source LMMs and the proprietary GPT-4V(ision) highlights the substantial challenges posed by MMMU. Even the advanced GPT-4V only achieves a 56% accuracy, indicating significant room for improvement. We believe MMMU will

stimulate the community to build next-generation multimodal foundation models towards expert artificial general intelligence.



# MMMU BENCHMARK

## Overview

We introduce the Massive Multi-discipline Multimodal Understanding and Reasoning (MMMU) benchmark, a novel benchmark meticulously curated to assess the expert-level multimodal understanding capability of foundation models across a broad scope of tasks. Covering subjects across disciplines, including Art, Business, Health & Medicine, Science, Humanities & Social Science, and Tech & Engineering, and over subfields. The detailed subject coverage and statistics are detailed in the figure. The questions in our benchmark were manually collected by a team of college students (including coauthors) from various disciplines and subjects, drawing from online sources, textbooks, and lecture materials.

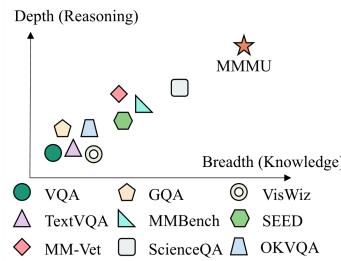
Art & Design	Business	Science
<p><b>Question:</b> Among the following harmonic intervals, which one is constructed incorrectly?</p> <p><b>Options:</b></p> <ul style="list-style-type: none"> <li>(A) Major third </li> <li>(B) Diminished fifth </li> <li><b>(C) Minor seventh </b></li> <li>(D) Diminished sixth </li> </ul> <p><b>Subject:</b> Music; <b>Subfield:</b> Music; <b>Image Type:</b> Sheet Music; <b>Difficulty:</b> Medium</p>	<p><b>Question:</b> ...The graph shown is compiled from data collected by Gallup &lt;image 1&gt;. Find the probability that the selected Emotional Health Index Score is between 80.5 and 82?</p> <p><b>Options:</b></p> <ul style="list-style-type: none"> <li>(A) 0</li> <li>(B) 0.2142</li> <li><b>(C) 0.3571</b></li> <li>(D) 0.5</li> </ul> <p></p> <p><b>Subject:</b> Marketing; <b>Subfield:</b> Market Research; <b>Image Type:</b> Plots and Charts; <b>Difficulty:</b> Medium</p>	<p><b>Question:</b> &lt;image 1&gt; The region bounded by the graph as shown above. Choose an integral expression that can be used to find the area of R.</p> <p><b>Options:</b></p> <ul style="list-style-type: none"> <li>(A) <math>\int_0^{1.5} [f(x) - g(x)] dx</math></li> <li>(B) <math>\int_0^{1.5} [g(x) - f(x)] dx</math></li> <li>(C) <math>\int_0^2 [f(x) - g(x)] dx</math></li> <li>(D) <math>\int_0^2 [g(x) - x(x)] dx</math></li> </ul> <p></p> <p><b>Subject:</b> Math; <b>Subfield:</b> Calculus; <b>Image Type:</b> Mathematical Notations; <b>Difficulty:</b> Easy</p>
<p><b>Question:</b> You are shown subtraction &lt;image 1&gt;, T2 weighted &lt;image 2&gt; and T1 weighted axial &lt;image 3&gt; from a screening breast MRI. What is the etiology of the finding in the left breast?</p> <p><b>Options:</b></p> <ul style="list-style-type: none"> <li>(A) Susceptibility artifact </li> <li>(B) Hematoma </li> <li><b>(C) Fat necrosis</b> </li> <li>(D) Silicone granuloma </li> </ul> <p><b>Subject:</b> Clinical Medicine; <b>Subfield:</b> Clinical Radiology; <b>Image Type:</b> Body Scans: MRI, CT.; <b>Difficulty:</b> Hard</p>	<p><b>Question:</b> In the political cartoon, the United States is seen as fulfilling which of the following roles? &lt;image 1&gt;</p> <p><b>Option:</b></p> <ul style="list-style-type: none"> <li>(A) Oppressor</li> <li>(B) Imperialist</li> <li><b>(C) Savior</b></li> <li>(D) Isolationist</li> </ul> <p></p> <p><b>Subject:</b> History; <b>Subfield:</b> Modern History; <b>Image Type:</b> Comics and Cartoons; <b>Difficulty:</b> Easy</p>	<p><b>Question:</b> Find the VCE for the circuit shown in &lt;image 1&gt;. Neglect VBE</p> <p>Answer: 3.75</p> <p>Explanation: ...IE = [(VEE) / (RE)] = [(5 V) / (4 k-ohm)] = 1.25 mA; VCE = VCC - IERL = 10 V - (1.25 mA) 5 k-ohm; VCE = 10 V - 6.25 V = 3.75 V</p> <p></p> <p><b>Subject:</b> Electronics; <b>Subfield:</b> Analog electronics; <b>Image Type:</b> Diagrams; <b>Difficulty:</b> Hard</p>

MMMU is designed to measure three essential skills in LMMs: perception, knowledge, and reasoning. Our aim is to evaluate how well these models can not only perceive and understand information across different modalities but also apply reasoning with subject-specific knowledge to derive the solution.

Our MMMU benchmark introduces key challenges to multimodal foundation models, as detailed in a figure. Among these, we particularly highlight the challenge stemming from the requirement for both expert-level visual perceptual abilities and deliberate reasoning with subject-specific knowledge. This challenge is vividly illustrated through our tasks, which not only demand the processing of various heterogeneous image types but also necessitate a model's adeptness in using domain-specific knowledge to deeply understand both the text and images and to reason. This goes significantly beyond basic visual perception, calling for an advanced approach that integrates advanced multimodal analysis with domain-specific knowledge.

## Comparisons with Existing Benchmarks

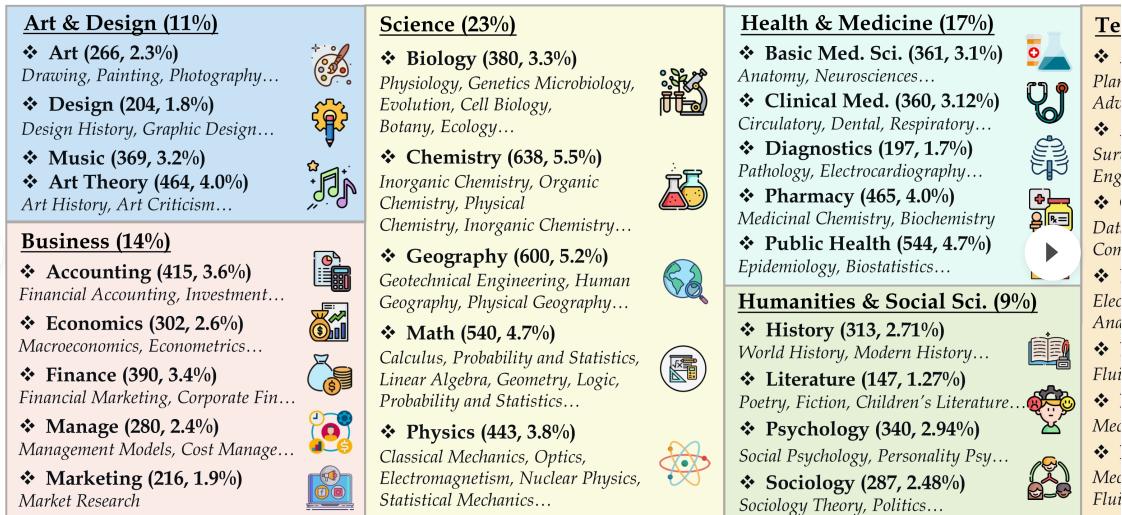
To further distinguish the difference between *dataset* and other existing ones, we elaborate the benchmark details in Figure. From the *breadth* perspective, the prior benchmarks are heavily focused on daily knowledge and common sense. The covered image format is also limited. Our benchmark aims to cover college-level knowledge with 30 image formats including diagrams, tables, charts, chemical structures, photos, paintings, geometric shapes, music sheets, medical images, etc. In the *depth* aspect, the previous benchmarks normally require commonsense knowledge or simple physical or temporal reasoning. In contrast, our benchmark requires deliberate reasoning with college-level subject knowledge.



Dataset	Size	Images	Format	Source	Answer
VQA	> 1M	V	I+T	Annotated	Open
GQA	> 1M	V	I+T	Synthesized	Open
VisWiz	32K	V	I+T	Annotated	Open
TextVQA	45K	OC	I+T	Annotated	MC
OKVQA	14K	V+OC	I+T	Annotated	Open
SEED	19K	V+OC	I+T	Annotated	MC
MMBench	3K	V+OC	I+T	Repurposed	MC
MM-Vet	0.2K	V+OC	I+T	Repurposed	MC
ScienceQA	6K	5 Types	I+T	Textbooks	MC
MMMU	11.5K	32 Types	Interleaved	Textbooks, Internet, Annotated	Open / MC

Sampled MMMU examples from each discipline. The questions and images need expert-level knowledge to understand and reason.

## Statistics



Sampled MMMU examples from each discipline. The questions and images need expert-level knowledge to understand and reason.

## EXPERIMENT RESULTS

### Leaderboard

We evaluate various models including LLMs and LMMs. In each type, we consider both closed- and open-source models. Our evaluation is conducted under a zero-shot setting to assess the capability of models to generate accurate answers without fine-tuning or few-shot demonstrations on our benchmark. For all models, we use the default prompt provided by each model for multi-choice or open QA, if available. If models do not provide prompts for task types in MMMU, we conduct prompt engineering on the validation set and use the most effective prompt for the later zero-shot experiment.

Click on MMMU-Pro, MMMU (Val) or MMMU (Test) to expand detailed results.

Reset		MMMU-Pro	MMMU(Val)	MMMU(Test)	
Name	Size	Date	Overall	Overall ↓	Overall
Human Expert (High)	-	2024-01-31	-	<b>88.6</b>	-
Human Expert (Medium)	-	2024-01-31	-	<u>82.6</u>	-
<b>o1</b>	-	2024-09-12	-	78.2*	-
Human Expert (Low)	-	2024-01-31	-	76.2	-
<b>Gemini 2.0 Pro</b>	-	2025-02-05	-	72.7*	-
<b>Gemini 2.0 Flash</b>	-	2025-02-05	-	71.7*	-
<b>Spark4.0-Turbo</b>	-	2025-01-22	-	70.4*	<u>60.3*</u>
<b>QVQ-72B-Preview</b>	72B	2024-12-25	-	70.3*	-
<b>InternVL2.5-78B</b>	78B	2024-12-05	48.6*	70.1*	<b>61.8*</b>
<b>GPT-4o (0513)</b>	-	2024-05-13	<b>51.9</b>	69.1*	-
<b>Claude 3.5 Sonnet</b>	-	2024-06-20	<u>51.5</u>	68.3	-
<b>Gemini 2.0 Flash-Lite</b>	-	2025-02-05	-	68.0*	-
<b>Gemini 1.5 Pro (0801)</b>	-	2024-08-01	46.9	65.8	-
<b>Qwen2-VL-72B</b>	72B	2024-08-29	46.2	64.5	-
<b>InternVL2.5-38B</b>	38B	2024-12-05	46.0*	63.9*	<b>57.6*</b>
<b>Gemini 1.5 Pro (0523)</b>	-	2024-05-23	43.5	62.2*	-
<b>InternVL2-Pro</b>	-	2024-07-04	-	62.0*	<b>55.7</b>
<b>TeleMM</b>	-	2024-11-18	-	61.4*	<b>58.2*</b>
<b>Llama 3.2 90B</b>	90B	2024-09-25	39.5*	60.3*	-
<b>NVLM-H 1.0 78B</b>	78B	2024-09-17	-	60.2*	<b>53.0*</b>
<b>InternVL2.5-26B</b>	26B	2024-12-05	37.1*	60.0*	<b>51.8*</b>
<b>NVLM-D 1.0 78B</b>	78B	2024-09-17	-	59.7*	<b>54.6*</b>
<b>Gemini 1.0 Ultra</b>	-	2023-12-06	-	59.4*	-
<b>Claude 3 Opus</b>	-	2024-03-05	-	59.4*	-
<b>GPT-4o mini</b>	-	2024-07-18	37.6	59.4*	-
<b>NVILA</b>	15B	2024-11-20	33.7*	58.7*	<b>51.8*</b>
<b>InternVL2-Llama3-76B</b>	76B	2024-07-04	40.0	58.3*	-
<b>RBDash V1.2</b>	72B	2024-08-21	-	57.7*	<b>51.2*</b>
<b>GPT-4V(ision) (Playground)</b>	-	2023-11-27	-	56.8	<b>56.1</b>
<b>LLaVA-OneVision-72B</b>	72B	2024-08-06	31.0	56.8*	-
<b>Reka Core</b>	-	2024-04-23	-	56.3*	-
<b>Gemini 1.5 Flash</b>	-	2024-05-23	-	56.1*	-
<b>InternVL2.5-8B</b>	8B	2024-12-05	34.3*	56.0*	<b>48.9*</b>
<b>InternVL2-40B</b>	40B	2024-07-04	34.2	55.2*	-
<b>Aria</b>	3.9B	2024-10-10	-	54.9*	-
<b>SenseChat-Vision-0423-Preview</b>	-	2024-04-23	-	54.6*	<b>50.3*</b>
<b>POINTS1.5</b>	7B	2024-12-11	-	53.8*	-
<b>Reka Flash</b>	-	2024-04-23	-	53.3*	-
<b>Claude 3 Sonnet</b>	-	2024-03-05	-	53.1*	-

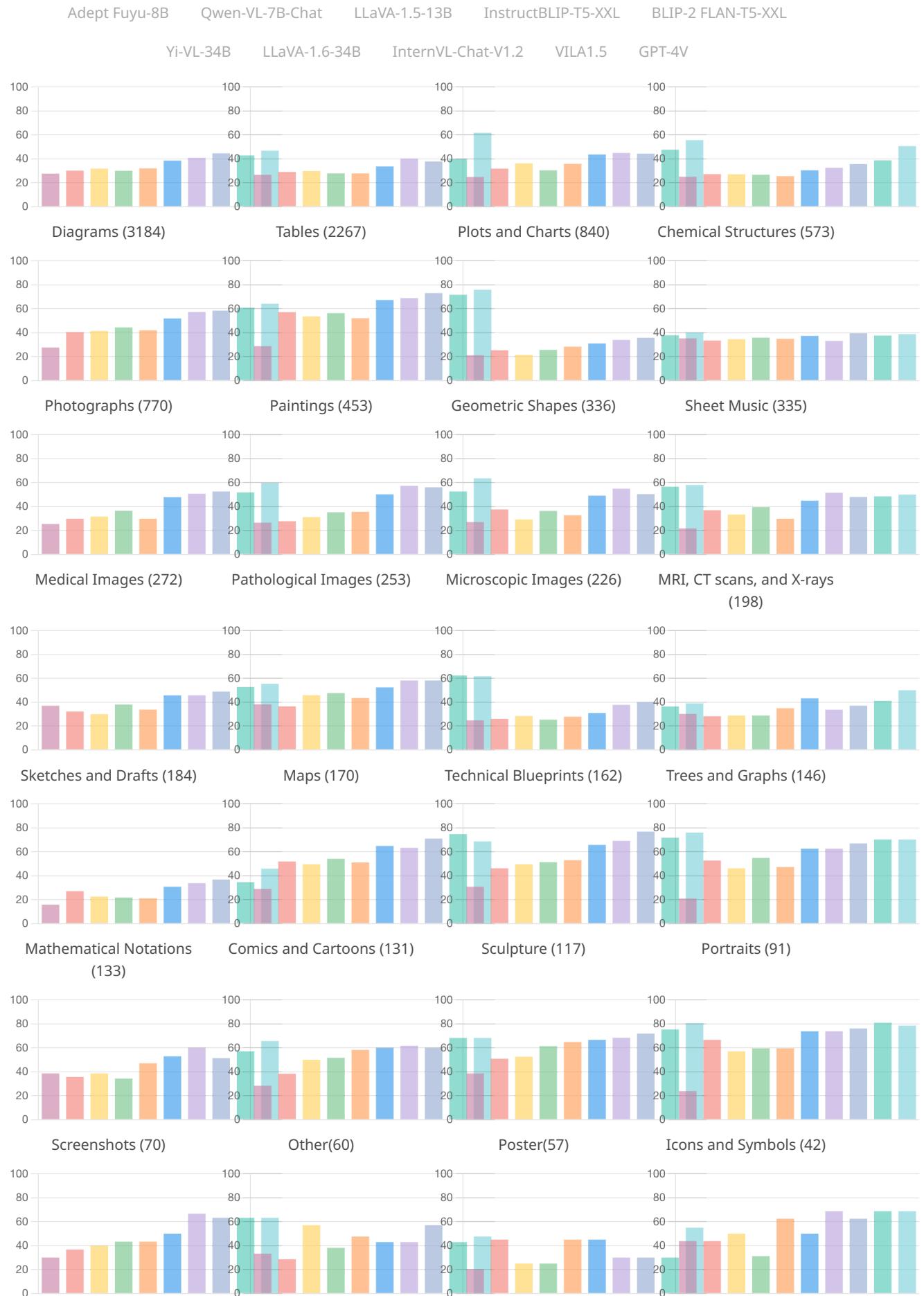
Reset		MMMU-Pro	MMMU(Val)	MMMU(Test)	
Name	Size	Date	Overall	Overall ↓	Overall
<a href="#">EVLM-KTO</a>	7B	2024-12-13	-*	53.0*	-
<a href="#">HPT Pro</a>	-	2024-03-16	-	52.0*	-
<a href="#">InternVL-Chat-V1.2</a>	-	2024-02-22	-	51.6*	46.2*
<a href="#">Qwen-VL-MAX</a>	-	2024-01-27	-	51.4*	46.8*
<a href="#">Skywork-VL</a>	-	2024-05-30	-	51.4*	46.2*
<a href="#">InternVL2-8B</a>	8B	2024-07-04	29.0	51.2*	-
<a href="#">DeepSeek-VL2</a>	4.5B	2024-12-13	-	51.1*	-
<a href="#">360VL</a>	70B	2024-05-21	-	50.8*	44.3*
<a href="#">MAmmoTH-VL-8B</a>	8B	2024-12-08	-*	50.8*	-
<a href="#">Llama 3.2 11B</a>	11B	2024-09-25	28.4*	50.7*	-
<a href="#">Claude 3 Haiku</a>	-	2024-03-05	-	50.2*	-
<a href="#">LlaVA-NEXT-72B</a>	72B	2024-01-30	25.1	49.9*	-
<a href="#">MiniCPM-V 2.6</a>	8B	2024-08-06	27.2	49.8*	-
<a href="#">Ovis1.5-Gemma2-9B</a>	9B	2024-06-14	-	49.8*	-
<a href="#">LLaVA-OneVision-7B</a>	7B	2024-08-06	24.1	48.8*	-
<a href="#">Ovis-Clip-Llama3-8B</a>	8B	2024-06-14	-	48.6*	39.0*
<a href="#">Adept Fuyu-Heavy</a>	-	2024-01-31	-	48.3*	-
<a href="#">LLaVA-NEXT-34B</a>	34B	2024-01-30	23.8	48.1*	44.7*
<a href="#">DeepSeek-VL2-Small</a>	2.8B	2024-12-13	-	48.0*	-
<a href="#">Gemini 1.0 Pro</a>	-	2023-12-06	-	47.9*	-
<a href="#">Idefics3-8B-Llama3</a>	8B	2024-08-22	22.9	46.6*	-
<a href="#">Marco-VL-Plus</a>	-	2024-03-09	-	46.2*	44.3*
<a href="#">Yi-VL-34B</a>	34B	2023-12-31	-	45.9*	41.5*
<a href="#">MiniCPM-Llama3-V 2.5</a>	8.5B	2024-05-20	19.6	45.8*	-
<a href="#">Qwen-VL-PLUS</a>	-	2023-12-15	-	45.2*	40.8*
<a href="#">HPT Air</a>	-	2024-03-16	-	44.0*	-
<a href="#">InternVL2.5-2B</a>	2B	2024-12-05	23.7*	43.6*	38.2*
<a href="#">Bunny-8B</a>	8B	2024-06-25	-	43.3*	39.0*
<a href="#">InternLM-XComposer2-VL</a>	-	2024-01-22	-	43.0*	38.2*
<a href="#">Phi-3.5-Vision</a>	-	2024-05-21	19.7	43.0*	-
<a href="#">Reka Edge</a>	-	2024-04-23	-	42.8*	-
<a href="#">MetaMorph</a>	-	2024-12-19	-	41.8*	-
<a href="#">HyViLM</a>	8B	2024-12-11	-	41.8*	-
<a href="#">Bunny-4B</a>	4B	2024-06-25	-	41.4*	38.4*
<a href="#">Marco-VL</a>	-	2024-01-16	-	41.2*	40.4*
<a href="#">OmniLMM-12B</a>	12B	2024-02-07	-	41.1*	37.8*
<a href="#">InternVL2.5-1B</a>	1B	2024-12-05	19.4*	40.9*	35.8*
<a href="#">DeepSeek-VL2-Tiny</a>	1.0B	2024-12-13	-	40.7*	-
<a href="#">InfiMM-Zephyr-7B</a>	7B	2024-01-18	-	39.4*	35.5*
<a href="#">Yi-VL-6B</a>	6B	2024-01-23	-	39.1*	37.8*
<a href="#">InternVL-Chat-V1.1</a>	-	2024-01-24	-	39.1*	35.3*

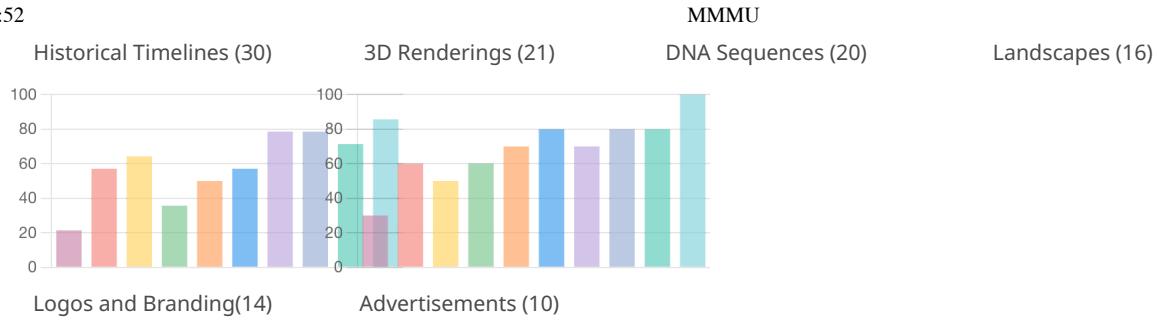
Reset		MMMU-Pro	MMMU(Val)	MMMU(Test)	
Name	Size	Date	Overall	Overall ↓	Overall
<b>Math-LLaVA-13B</b>	13B	2024-06-25	-	38.3*	34.6*
<b>Bunny-3B</b>	3B	2024-02-13	-	38.2*	33.0*
<b>LLaVA-UHD v2</b>	-	2024-12-21	-	38.2*	-
<b>ILLUME</b>	-	2024-12-09	-	38.2*	-
<b>SVIT</b>	-	2023-12-26	-	38.0*	34.1*
<b>MiniCPM-V</b>	-	2024-02-01	-	37.2*	-
<b>MiniCPM-V 2.0</b>	-	2024-04-12	-	37.1*	-
<b>LLaVA-1.5-13B</b>	13B	2023-11-27	-	36.4	33.6
<b>Emu2-Chat</b>	-	2023-12-24	-	36.3*	34.1*
<b>LLaVA-NeXT-13B</b>	13B	2024-01-30	17.2	36.2*	-
<b>Qwen-VL-7B-Chat</b>	-	2023-11-27	-	35.9	32.9
<b>LLaVA-NeXT-Vicuna-7B</b>	7B	2024-01-30	16.1	35.8*	-
<b>InstructBLIP-T5-XXL</b>	-	2023-11-27	-	35.7	33.8
<b>BLIP-2 FLAN-T5-XXL</b>	-	2023-11-27	-	35.4	34.0
<b>LLaVA-NeXT-mistral-7B</b>	7B	2024-01-30	17.0	35.3*	-
<b>BLIP-2 FLAN-T5-XL</b>	-	2023-11-27	-	34.4	31.0
<b>SynerGen-VL</b>	2.4B	2024-12-12	-	34.2*	-
<b>HoVLE</b>	2.6B	2024-12-18	-	33.7*	-
<b>HoVLE (HD)</b>	2.6B	2024-12-18	-	33.2*	-
<b>InstructBLIP-T5-XL</b>	-	2023-11-27	-	32.9	30.6
<b>SPHINX</b>	-	2023-12-26	-	32.9*	32.9*
<b>Olympus</b>	2.7B	2024-12-12	-	32.8*	-
<b>mPLUG-OWL2</b>	-	2023-12-04	-	32.7*	32.1*
<b>Gemini Nano2</b>	-	2023-12-06	-	32.6*	-
<b>Otter</b>	-	2023-11-27	-	32.2	29.1
<b>CogVLM</b>	-	2023-11-27	-	32.1	30.1
<b>Emu3</b>	8B	2024-09-27	-	31.6*	-
<b>LLaMA-Adapter2-7B</b>	7B	2023-11-27	-	29.8	27.7
<b>OpenFlamingo2-9B</b>	9B	2023-11-27	-	28.7	26.3
<b>Adept Fuyu-8B</b>	8B	2023-11-27	-	27.9	27.3
<b>Show-o</b>	1.3B	2024-08-22	-	27.4*	-
<b>MiniGPT4-Vicuna-13B</b>	13B	2023-11-27	-	26.8	27.6
<b>Frequent Choice</b>	-	2023-11-27	12.1	26.8	23.9
<b>Kosmos2</b>	-	2023-11-27	-	24.4	26.6
<b>Random Choice</b>	-	2023-11-27	12.6	22.1	25.8
<b>Weitu-VL-1.0-15B</b>	15B	2024-04-01	-	-	38.4*

Overall results of different models on the MMMU leaderboard. The best-performing model in each category is **in-bold**, and the second best is underlined. \*: results provided by the authors.

## Different Image Types

We compare the performance of various models across top frequent image types. Across all types, GPT-4V consistently outperforms the other models by a huge margin. Open-source models demonstrate relatively strong performance in categories like Photos and Paintings, which are more frequently seen during training. However, for less common image categories like Geometric shapes, Music sheets and Chemical structures, all models obtain very low scores (some are close to random guesses). This indicates that the existing models are generalizing poorly towards these image types.



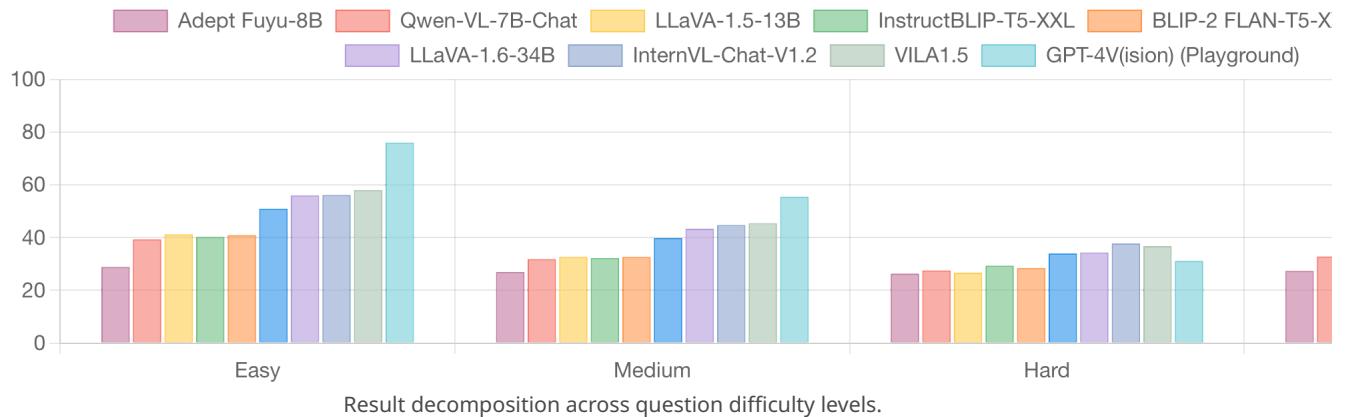


Selected models' performance on 30 different image types. Note that a single image may have multiple image types.

## Different Difficulty Levels

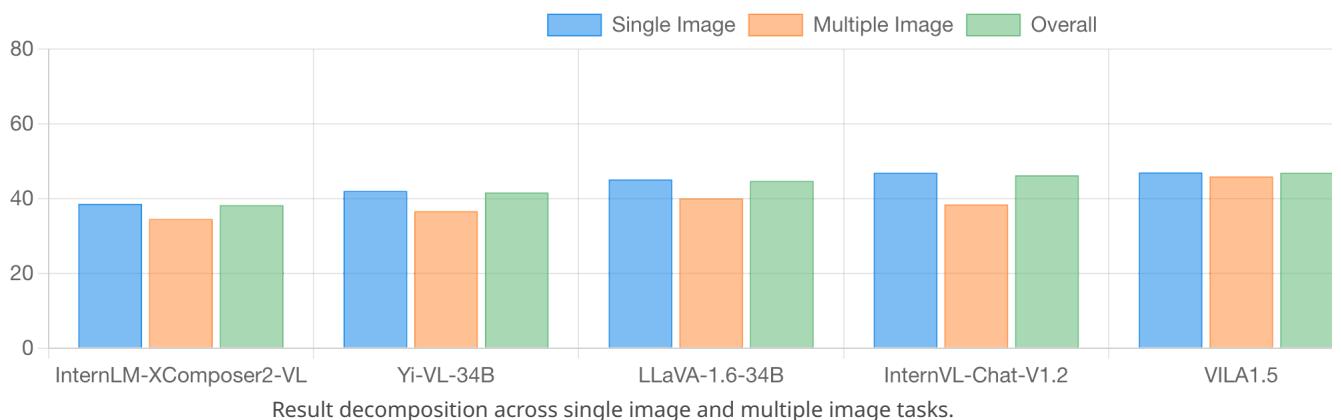
We compare the performance of selected models across three difficulty levels. GPT-4V demonstrates a significantly higher proficiency, with a success rate of 76.1%, compared to open-source models in the "Easy" category. When it comes to the "Medium" category, while the gap narrows, GPT-4V still leads at 55.6%. The further diminishing performance gap in the "Hard" category across models indicates that as the complexity of tasks increases, the advantage of more advanced models like GPT-4V almost disappears. This might reflect a current limitation in handling expert-level challenging queries even for the most advanced models.

Click legend to switch the comparison chart.



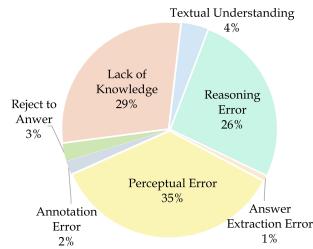
## Single Image VS Multiple Image

Click legend to switch the comparison chart.



## Error Analysis

We delve into the analysis of errors by GPT-4V, a pivotal aspect for understanding its operational capabilities and limitations. This analysis serves not only to identify the model's current shortcomings but also to guide future enhancements in its design and training. We meticulously examine 150 randomly sampled error instances from GPT-4V's predictions. These instances are analyzed by expert annotators who identify the root causes of mispredictions based on their knowledge and the golden explanations if available. The distribution of these errors is illustrated in Figure, and a selection of 100 notable cases, along with detailed analyses, is included in the Appendix.



Error distribution over 150 annotated GPT-4V errors.

## Error Examples

X
Accounting: Financial Accounting

**Question:** From the following Company Y adjusted trial balance, what is the retained earnings to be reported? <image 1>

**Option:**

(A) \$17,800  
 (B) \$16,800  
 (C) \$15,550  
 (D) \$12,280

Adjusted Trial Balance		
	Debit	Credit
Cash	\$ 32,000	
Accounts receivable	17,300	
Prepaid insurance	6,400	
Land	10,000	
Accounts payable		\$ 10,900
Salaries payable		6,000
Common stock		31,000
Retained earnings		4,200
Dividends	8,000	
Service revenue		74,000
Insurance expense	5,600	
Salaries expense	24,000	
Miscellaneous expense	22,800	
	126,100	126,100

<image 1>

**GPT-4V:** To calculate the retained earnings to be reported for Company Y from the adjusted trial balance, we need to consider the following components:

1. Beginning Retained Earnings: This would be the retained earnings balance at the start of the period. However, since it is not provided, we can assume it is included in the current retained earnings figure in the trial balance or that this is the first year of operation, and as such, it would be zero.
2. Add: Net Income (or Less: Net Loss): Net income is calculated as revenues minus expenses. If expenses exceed revenues, it results in a net loss.
3. Less: Dividends: These are distributions to shareholders, which reduce retained earnings. From the adjusted trial balance provided, the calculation would be as follows:

Net Income: Service Revenue (\$74,000) - (Insurance Expense \$5,600 + Salaries Expense \$24,000 + Miscellaneous Expense \$22,800) = \$74,000 - \$52,400 = \$21,600.

Less Dividends: \$8,000. Retained Earnings to be Reported = Beginning Retained Earnings (if given, assumed to be included in the trial balance) + Net Income - Dividends. The trial balance shows Retained Earnings as \$6,000. Assuming this is the beginning balance and adding the net income less dividends: Retained Earnings to be Reported = \$6,000 + \$21,600 - \$8,000 = \$19,600. There is no answer matching the option.

**Error Category: Perceptual Error**

**Error Reason:** GPT-4V understands the question and can know how to calculate the Retained Earnings to be Reported. Most of the calculation process is correct. However, in the final step, GPT-4V wrongly extracts the number of Retained Earnings from the table image (should be \$4, instead of \$6,000).

**Ground Truth: (A) \$17,800**

<https://mmmu-benchmark.github.io>

9/11

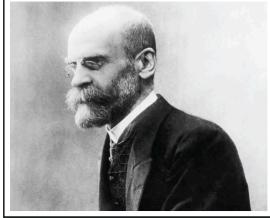
## Correct Examples

 **Sociology: Sociology Theory**

**Question:** For the person in <image 1>, what is the root of everything social?

**Option:**

- (A) Collective Effervescence
- (B) Society
- (C) The Economy
- (D) Religion



<image 1>

**GPT-4V:** For Émile Durkheim, one of the founding figures in sociology, the root of everything social could be argued to be "Collective Effervescence," a term he coined. This concept refers to moments of communal passion and ecstasy that occur during group rituals and events, which Durkheim believed were the basis for the solidarity of social groups and ultimately for society itself. It's a foundational element in his theory on the formation of religions and societies.

The correct answer to the question, considering Emile Durkheim's sociological theories, is "Religion." Durkheim viewed religion as a fundamental social institution that is central to the formation and sustenance of societies. He believed that religious beliefs and rituals play a critical role in creating social cohesion and collective conscience.

**Ground Truth: (D) Religion**

**Explanation:** Durkheim believed that religion is the root of everything social. He argued that religion provides the framework and moral values that bind individuals together and create social cohesion. Through religious rituals and beliefs, people experience a sense of collective effervescence, a heightened state of group solidarity and shared emotions. This collective effervescence reinforces social norms, strengthens social bonds, and gives individuals a sense of belonging to a larger community. Therefore, Durkheim viewed religion as the foundation of social order and the driving force behind societal cohesion.

## BibTeX

```
@inproceedings{yue2023mmmu,
  title={MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI},
  author={Xiang Yue and Yuansheng Ni and Kai Zhang and Tianyu Zheng and Ruqi Liu and Ge Zhang and Samuel Stevens},
  booktitle={Proceedings of CVPR},
  year={2024},
}
```

This website is website adapted from [Nerfies](#) and [MathVista](#), licensed under a [Creative Commons Attribution-ShareAlike 4.0 International License](#).